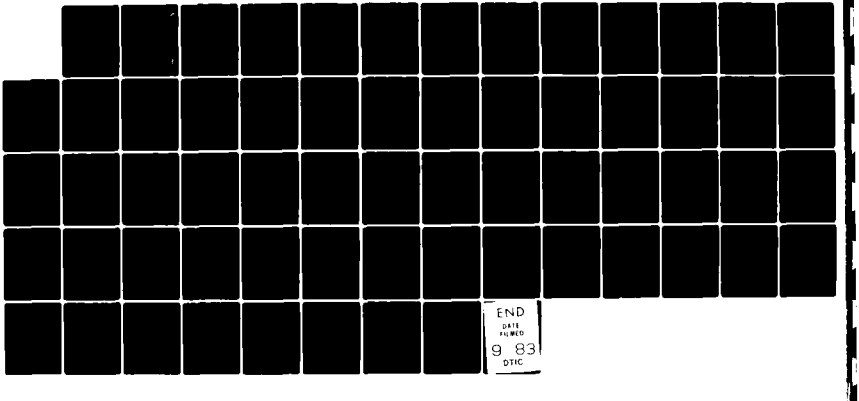
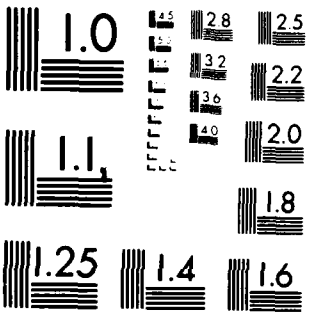
AD-A131 394 NONPARAMETRIC ESTIMATION BY THE METHOD OF SIEVES(U) 1/1
BROWN UNIV PROVIDENCE RI DIV OF APPLIED MATHEMATICS
S GEMAN ET AL. JUL 83 RPA-131 ARO-17444.4-MA
UNCLASSIFIED DAAG29-80-K-0006 F/G 12/1 NL





NONPARAMETRIC ESTIMATION BY THE METHOD OF SIEVES

FINAL TECHNICAL REPORT

Stuart Geman
Donald E. McClure

July 1983

U. S. ARMY RESEARCH OFFICE

Contract DAAG29-80-K-0006

Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

Reports in Pattern Analysis No. 131

THE VIEWS, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHORS AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A137 394	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Nonparametric Estimation by the Method of Sieves		5. TYPE OF REPORT & PERIOD COVERED Final Technical 1 June 1980 - 31 May 1983
		6. PERFORMING ORG. REPORT NUMBER RPA No. 131
7. AUTHOR(s) Stuart Geman Donald E. McClure		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-K-0006
9. PERFORMING ORGANIZATION NAME AND ADDRESS Division of Applied Mathematics Brown University Providence, R.I. 02912		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE July 1983
		13. NUMBER OF PAGES 61
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Nonparametric statistics, estimation, regression, method of sieves, maximum likelihood, density estimation, kernel estimators, cross-validation, ridge regression, Poisson process, intensity function estimation, Radon transform, image processing, surface reconstruction, computed tomography, image registration, consistency, asymptotic distribution.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
The research project has built a theoretical foundation for using the method of sieves to adapt classical estimation principles such as maximum likelihood and least squares to problems with infinite dimensional parameter spaces. The first results about consistency of cross-validated estimators of density functions have been obtained. The method of sieves and the principle of maximum likelihood have been used to develop algorithms for digital image processing. Specific applications include image segmentation,

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (abstract cont'd.)

reconstruction methods for tomography, image registration methods for moving objects, and surface restoration algorithms.

I. Introduction

Since the June 1, 1980 starting date of our contract with the U. S. Army Research Office (DAAG29-80-K-0006), significant progress has been made in the areas of proposed research. We believe that the power and versatility of Grenander's Method of Sieves as a general approach to nonparametric estimation is now well established. Depending on the chosen sieve, it offers an array of new approaches in specific applications and it frequently contributes to our understanding of well-studied estimators and suggests improvements of them. Although our theoretical understanding is not complete, we now have a firm mathematical and intuitive understanding of the method. In the past year and a half we have been increasingly supplementing our analytic approach with more detailed explorations of specific practical applications.

In Section II we shall review the essential ideas behind the method of sieves and lay the foundations of the problems that have guided our research during the contract period. Section III summarizes our progress and highlights concrete results that have been obtained. The summary identifies four broad areas where progress has been made: the general theory of sieves for nonparametric estimation, theoretical and methodological aspects of selecting sieve parameters (cross-validation), studies of specific method-of-sieve estimators, and methodology and development of reources for computer experiments. Section IV describes projects that are currently being pursued, including manuscripts in various stages of preparation and collaborative efforts on applications of

nonparametric estimation theory and the method of sieves to digital picture processing. Image processing problems of segmentation, registration, and reconstruction have proven to be a fertile source of interesting and important theoretical and applied problems in nonparametric estimation. Section V lists the technical reports and papers that document progress on specific research problems in our project. Section VI acknowledges the personnel who have contributed to the effort.

II. The Method of Sieves[1]

Techniques for estimating finite dimensional parameters typically fail when applied to infinite dimensional problems. The difficulties encountered in moving from finite to infinite dimensions are well illustrated by the failure of maximum likelihood in nonparametric density estimation. Let $x_1,\ldots,x_n$ be an iid sample from an absolutely continuous distribution with unknown probability density function (pdf) $\alpha_0(x)$. The maximum likelihood estimator for α_0 maximizes

$$\prod_{i=1}^{n} \alpha(x_i) \tag{1}$$

over some specified set of candidates. But if this set is too large, then the method will fail to produce a meaningful estimator.

[1] Much of the material in this section appeared in a paper by Geman and Hwang [19]. It is included here for the purpose of making this proposal, as nearly as possible, self-contained.

For instance, in the extreme case nothing is known about α_0, and the maximum of (1) is not achieved. Roughly speaking, we move out of the parameter space (the space of all densities) toward a discrete distribution with jumps at the sample points.

Another example of the failure of classical methods to solve infinite dimensional problems is the breakdown of least squares in the nonparametric estimation of a regression. Let X and Y be random variables and let $(x_1,y_1),\ldots,(x_n,y_n)$ be an iid sample from the bivariate distribution of (X,Y). The least squares estimator of the regression function $E(Y|X=x)$ minimizes

$$\sum_{i=1}^{n}\{y_i-\alpha(x_i)\}^2. \tag{2}$$

Observe that the minimum is zero and is achieved by any function which passes through all of the points of observation $(x_1,y_1),\ldots,(x_n,y_n)$. Excepting some very special cases, this set does not in any meaningful sense converge to the true regression.

Grenander [24] suggests the following remedy: perform the optimization (maximization of the likelihood, minimization of the sum of square errors, etc.) within a subset of the parameter space, and then allow this subset to "grow" with the sample size. He calls this collection of subsets from which the estimator is drawn a "sieve," and the resulting estimation procedure is his "method of sieves." The method leads easily to consistent nonparametric estimators in even the most general settings, with different sieves giving rise to different estimators. Often the sieve estimator is closely related to an already well-studied

estimator, and may suggest an improvement, or a new point of view and a new motivation.

The histogram is a simple example of a sieve estimator. Consider again the problem of estimating an entirely unknown density function $\alpha_0(x)$. We have seen that unmodified maximum likelihood is not consistent for this problem. A sieve is an indexed collection of subsets of the parameter space, such as:

$$S_\lambda = \{\alpha: \text{ is a pdf which is constant on } [\tfrac{k-1}{\lambda},\tfrac{k}{\lambda}), \quad k=0,\pm1,\pm2,\ldots\}$$

$\lambda > 0$. The method of sieves estimator maximizes the likelihood $\Pi_{i=1}^n \alpha(x_i)$, subject to $\alpha \in S_\lambda$, allowing λ to grow with the sample size. The well-known solution is the function

$$\hat{\alpha}(x) = \frac{\lambda}{n} \#\{x_i: \tfrac{k-1}{\lambda} \leq x_i < \tfrac{k}{\lambda}\} \text{ for } x \in [\tfrac{k-1}{\lambda},\tfrac{k}{\lambda}),$$

i.e. the histogram with bin width λ^{-1}. Putting aside details, we know that if $\lambda_n \uparrow \infty$ sufficiently slowly, then $\hat{\alpha}$ is consistent, e.g. in the sense that $\int |\hat{\alpha}(x)-\alpha_0(x)|dx \to 0$ a.s.

For the same problem, a different and more interesting sieve is the "convolution sieve":

$$S_\lambda = \{\alpha:\alpha(x) = \int \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}(x-y)^2} F(dy), \text{ F an arbitrary cdf}\}.$$

This time, maximizing the likelihood within S_λ gives rise to an estimator closely related (but not identical) to the Parzen-Rosenblatt (Gaussian) kernel estimator. In fact, the latter is in the sieve S_λ: take F to be the empirical distribution

function. But the maximum of the likelihood is achieved by using a different distribution. As with the Parzen-Rosenblatt estimator, if $\lambda_n \uparrow \infty$ sufficiently slowly (i.e. the "window width" is decreased sufficiently slowly) then the estimator is consistent. A more precise discussion of this and some related sieves is in section III.

The inconsistency of least squares nonparametric regression can be similarly rectified by introducing sieves. Let us look again at the regression problem formulated above; recall that $(x_1,y_1),\ldots,(x_n,y_n)$ is an iid sample from the bivariate distribution of (X,Y). Given a sieve S_λ, the method of sieves estimator $\hat{\alpha}$ minimizes the sum of square errors, (2), subject to $\alpha \in S_\lambda$. If, as an example,

$$S_\lambda = \{\alpha:\alpha \text{ absolutely continuous, } \int |\frac{d}{dx}\alpha(x)|^2 dx \leq \lambda\},$$

then $\hat{\alpha}$ is uniquely determined; it is a first degree polynomial smoothing spline; i.e. $\hat{\alpha}$ is continuous and piecewise linear with discontinuities in $(d/dx)\hat{\alpha}$ at $x_1,\ldots,x_n$; see Schoenberg [41]. It is possible to show that if λ_n increases sufficiently slowly, then the estimator is strongly consistent for $E(Y|X=x)$ in a suitable metric; details are in Geman [18]. Other sieves applied to the same problem lead to kernel estimators and still others to new estimators. Even if the squared loss function $\{y-\alpha(x)\}^2$ is replaced by a "robust" alternative, minimization over too large a set will again fail to produce a meaningful estimator. In exactly the same way, sieves offer a remedy in this case as well.

III. Progress During Current Contract Period

We will describe in detail our progress during the current contract period. Most of what we will say is included in finished, and in some cases already published, manuscripts. In addition to making our presentation self-contained, the details included here will make our later discussion of proposed research much clearer.

A. A General Consistency Argument for Method of Sieves Estimators.

We began our work on the method of sieves by formulating theorems establishing the consistency of sieve estimators derived from the maximum likelihood procedure. For the problems addressed, these theorems hold in the most general possible settings. In our original proposal we expressed our intention to extend these theorems to apply to a wider variety of problems, including other criterion functions (such as least squares, robust loss functions, etc.) and including observations, such as continuous time stochastic processes, which are not naturally indexed by a discrete parameter. We found, however, that this greater scope led to theorems whose application required the verification of numerous complicated conditions.

The consistency problem for the method of sieves is essentially one of identifying an appropriate upper bound on the rate of growth for the sieve. Instead of a general theorem for this identification, we have developed a general approach. What we have is a versatile strategy for determining bounds on

growth rates of sieves, and it easily achieves all of the desired extensions. Whereas a rigorous explanation of the technique would be unnecessarily involved, we can briefly describe it in a heuristic manner and refer to several papers for details of its application.

Our approach to establishing consistency for sieves is an adaptation of methods first developed by Bahadur [5], Wald [56], and others in similar connection with extensions of the maximum likelihood procedure. It might be called the "small-ball technique", as it consists of partitioning the sieve into balls sufficiently small so that all estimators within each ball behave similarly. One first demonstrates that when using the estimator at the center of one of these balls the value of the criterion function (log-likelihood, sum of square errors, etc.) is well-approximated by its expected value, at least with high probability. If this approximation can be made to hold uniformly over all centers, then it will also hold over the entire sieve - assuming, again, that the balls are sufficiently small. As the sieve grows, the balls are made smaller and more numerous. If the sieve growth is sufficiently slow then a "uniform law of large numbers" can be established: optimizing the criterion function over the sieve is asymptotically equivalent to optimizing its expected value. The point is that the latter optimization typically defines the target parameter. As an example let α_0 be an unknown density function, and consider the likelihood crtierion: the maximization of

$$\frac{1}{n} \sum_{i=1}^{n} \log \alpha(x_i)$$

over $\alpha \in S_\lambda$ is asymptotically equivalent to the maximization of

$$E \frac{1}{n} \sum_{i=1}^{n} \log \alpha(x_i) = \int \alpha_0(x) \log \alpha(x) dx$$

provided the sieve growth is sufficiently slow (λ grows slowly with n). It is easily demonstrated that

$$\int \alpha_0(x) \log \alpha(x) dx$$

achieves its maximum at $\alpha = \alpha_0$, the unknown density. Nonparametric regression by least squares leads to the problem of minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha(x_i))^2$$

over a sieve $\alpha \in S_\lambda$ (we use the notation introduced in section II). Provided that the sieve growth is sufficiently slow, this is asymptotically equivalent to the minimization of

$$E \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha(x_i))^2 = E(Y - \alpha(X))^2.$$

For the latter, a minimum is achieved by the regression $\alpha(x) = E\{Y|X=x\}$, i.e. by the target parameter. For robust loss functions, the minimizer of the expected criterion function may or may not be the target parameter, depending on the true underlying distribution. Hence, sieve estimators derived from robust loss functions have an added condition for consistency.

By now we have numerous examples of the application of this technique to problems of density estimation, regression (with and without robust loss functions), the estimation of convex

sets from noisy one dimensional projections (a problem arising in emission tomography), the estimation of images of conformal mappings from noisy boundary observations, and the estimation of the mean and drift functions for various stochastic processes. For the latter, we are able to exend our original results from the artificial restriction of using only partial and discrete observations of a continuous time process to the more natural and no doubt more efficient use of the entire continuous time development of the process. Details on some of these applications of the "small-ball technique" can be found in the references [9],[17],[18],[19], and [35].

B. Cross-validated Smoothing.

In our original proposal we indicated that consistency results for estimators smoothed by cross-validation would be given the highest priority in our research program. In this section, we will review the role of cross-validation in nonparametric estimation, and we will summarize our progress in establishing properties of estimators smoothed by this technique.

1. The Smoothing Problem

The practical application of nonparametric (infinite dimensional) estimators requires the choice of a "smoothing parameter". For estimators generated by the Method of Sieves, we can specify as a function of sample size an upper bound on the rate of growth of the sieve which will guarantee consistency, but this rate tells us very little about the practical choice of a sieve size when faced with a finite fixed sample. Too

large a sieve produces an undersmoothed, or "overfit", estimator, and too small a sieve produces an estimator that is little influenced by the observations - an oversmoothed estimator. The proper choice of sieve size for finite fixed samples is the smoothing problem for the method of sieves, and it has its analogue for all non-Bayesian estimators of infinite dimensional parameters. Thus the histogram requires the choice of a bin-width; the kernel estimator requires the choice of a window width; the maximum penalized likelihood estimator ([22], [51]) requires that a weight be assigned the penalty term; and orthogonal series estimators must be suitably truncated [29], or "band limited" [54]. Regarding kernel estimators, Silverman [46] observes that "there seems to be considerable need for objective methods of determining the window width appropriate to a given sample". Speaking more generally, Wahba [54] remarks: "A major problem in density estimation is to choose the smoothing parameter(s), which are part of every density estimate...". And the problem is not peculiar to density estimation. Splines, kernels, and the newer "recursive partitions" (see, for example, [23]) for nonparametric regression, all require first a version of smoothing to be fully defined.

To illustrate the problem more concretely, in the context of the method of sieves, let us return to the convolution sieve for density estimation introduced earlier in section II:

$$S_\lambda = \{\alpha : \alpha(x) = \int \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}(x-y)^2} F(dy),\ F \text{ an arbitrary cdf}\}.$$

(This example also gives us an opportunity to present some additional work done during the current contract period.) Recall our description of the associated maximum likelihood estimator as being closely related to the Parzen-Rosenblatt kernel estimator. More specifically (see [20] for details):

Proposition[1] For each n and λ define

$$M_\lambda^n = \{\alpha \in S_\lambda : \prod_{i=1}^n \alpha(x_i) \sup_{\beta \in S_\lambda} \prod_{i=1}^n \beta(x_i)\}$$

= set of maximum likelihood solutions. For every n and λ M_λ^n is nonempty, and if $\alpha \in M_\lambda^n$ then

$$\alpha(x) = \sum_{i=1}^n p_i \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}(x-y_i)^2}$$

for some $y_1,\ldots,y_n$ and $p_1,\ldots p_n$ satisfying $p_i \geq 0$, $1 \leq i \leq n$, $\sum_{i=1}^n p_i = 1$. Furthermore, if $\min(x_1,\ldots x_n) < \max(x_1,\ldots x_n)$ then $\min(x_1,\ldots x_n) < \min(y_1,\ldots y_n)$ and $\max(y_1,\ldots y_n) < \max(x_1,\ldots x_n)$.

It is interesting to note that the kernel estimator (with Gaussian kernel)

$$\beta(x) = \frac{1}{n} \sum_{i=1}^n \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}(x-x_i)^2}$$

is in S_λ, but the last statement in the proposition indicates that β is not among the maximum likelihood solutions, i.e.

---

[1] If in the definition of the sieve S_λ the Gaussian kernel is replaced by a double exponential, then a more complete characterization of the maximum likelihood solution is available. This characterization was presented recently by Blum and Walter [6] in a paper that also makes an interesting connection between the convolution sieve and other methods for nonparametric density estimation.

$\beta \notin M_\lambda^n$.[1]

Although we have characgerized the maximum likelihood set up to the 2n parameters $y_1,\ldots y_n$, $p_1,\ldots p_n$, its actual computation is difficult. The proposition suggests a smaller and computationally more attractive sieve:

$$\hat{S}_\lambda = \{\alpha:\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}(x-y_i)^2}\}$$

i.e. we give equal mass to each kernel, but allow the locations to move in such a way as to maximize the likelihood. (Here again, it can be shown that the kernel estimator is not among the maximum likelihood solutions.) We have experimented with $\hat{S}_\lambda$ and have found, as a rule, that the number of distinct y's in a maximum likelihood solution is considerably smaller than n. In other words, the kernels will often coalesce to achieve an increased likelihood. Sometimes this results in strikingly accurate density estimators, while at other times this "maximum likelihood" solution is a poor second to the corresponding (same window width) kernel estimator. In either case, this estimator suffers the very same stability problem as the kernel estimator: the results are critically dependent on the choice of the kernel width (which is here governed by the sieve parameter λ). The important choice of a good λ is the smoothing problem for this estimator.

---

[1] Again referring to the paper by Blum and Walter, the surprising result for double exponential kernels is that the kernels are centered at the observations. However, the classic kernel estimator is still, in general, not the maximum likelihood solution.

One approach to this critical dependence on window width (call it σ) is to include σ as a free parameter within the sieve, and thus allow it to be chosen by maximum likelihood. But we must be somewhat careful; we cannot merely replace $\hat{S}_\lambda$ by

$$\{\alpha:\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-y_i)^2} \},$$

since then the maximum of the likelihood is achieved with $\sigma=0$ and the kernels centered at the sample points. Let us instead define the sieve parameter λ to be the number of kernels, restricting this to be smaller than n ($\lambda < n$), and consider

$$\tilde{S}_\lambda = \{\alpha:\alpha(x) = \frac{1}{\lambda}\sum_{i=1}^{\lambda} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-y_i)^2} \}.$$

The associated maximum likelihood estimator has performed well in our simulations, and we can demonstrate (by an application of the "small-ball technique") that if $\lambda_n \to \infty$ in such a way that $\lambda_n = 0(n^{1/5-\varepsilon})$ for some $\varepsilon > 0$ then

$$\sup_{\alpha\in\tilde{M}^n_{\lambda_n}} \int|\alpha(x)-\alpha_0(x)|dx \to 0 \quad \text{a.s.}$$

where $\tilde{M}^n_\lambda$ is the maximum likelihood set associated with the sieve S_λ, and α_0 is the true density. But for fixed sample the degree of smoothing is still important: with moderate sample sizes ($n \approx 50$), $\lambda=1$ generally oversmooths and $\lambda=n-1$ will almost always drastically undersmooth.

Among the important practical questions about sieves that remain unanswered (including relative efficiency, asymptotic distributions, good sieves for robust estimation, etc.), perhaps most pressing is this smoothing problem. Probably the most broadly applicable and widely studied of the general solutions proposed to the smoothing problem is the method of cross-validation (for an introduction, see Stone [50] and Wahba [54] and the many references therein). We have experimented extensively with the method and have found what many others have found (see [44],[53],[54], and [55] to name just a few): that cross-validation is often times a strikingly effective means of choosing an appropriate degree of smoothing. But success with the method is not guaranteed. An example of its failure on a seemingly canonical problem is due to Schuster and Gregory [42]. They prove that cross-validated kernel estimators with compact kernels are inconsistent for the exponential distribution. They also demonstrate by simulation that these estimators have poor small sample behavior. Other authors have raised doubts about the effectiveness of cross-validation for choosing the smoothing parameter in ridge regression (see, e.g. [14],[30]).

Since there is very little known analytically about cross-validated estimators, and since there seems to be a real practical need for a better understanding of the method (especially in identifying problems to which it can be successfully applied), and since the method offers a natural solution to the smoothing problem for the Method of Sieves, we have spent considerable time

in a mathematical study of cross-validation. Subsection 2 below is a brief introduction to the method, and subsection 3 presents the analytic results about cross-validation that we have obtained during the current contract period.

2. Smoothing by Cross-validation

The general idea behind data-driven smoothing is to measure, as a function of the smoothing parameter (call it λ), the ability of the estimator to "explain", or to "fit", the observed data. λ is chosen to maximize this measure of explanation. When smoothing by cross-validation, in particular, the measure of explanation is obtained by successively deleting single observations, computing the estimator from the remaining observations, and then applying the estimator to the deleted observation. The details are most easily illustrated with specific examples. For this purpose, we will use the kernel estimator, call if $f_{\lambda,n}$:

$$f_{\lambda,n}(x) = \frac{1}{n} \sum_{j=1}^{n} \lambda K(\lambda(x-x_j))$$

for some probability kernel K. We will denote by $f^i_{\lambda,n-1}$ the estimator computed after deleting the i'th observation, i.e.

$$f^i_{\lambda,n-1}(x_j) = \frac{1}{n-1} \sum_{j\neq i} \lambda K(\lambda(x-x_j)).$$

Now $f^i_{\lambda,n-1}$ is not dependent on x_i, and $f^i_{\lambda,n-1}(x_i)$ may be taken as a measure of the appropriateness of λ as a value for the smoothing parameter: If $f^i_{\lambda,n-1}(x_i)$ is large, then it might be

said that $f^i_{\lambda,n-1}$ "anticipated" the observation x_i, and that λ is an appropriate degree of smoothing (at least for samples of size n-1); small values of $f^i_{\lambda,n-1}(x_i)$ suggest that the observation x_i was unlikely (under the density $f^i_{\lambda,n-1}$), and may be interpreted as evidence against the appropriateness of λ. As i ranges through the full sample we obtain n such measures of fit, and these may be combined into a likelihood-like expression:

$$(3) \qquad L_\lambda = \prod_{i=1}^{n} f^i_{\lambda,n-1}(x_i).$$

One version of cross-validated density estimation (first proposed by Habbema et al. [26], and separately by Duin [12]) chooses λ to maximize L_λ (call this value λ^*, the "cross-validated smoothing parameter"), and then forms the corresponding estimator, $f_{\lambda^*,n}$ (the "cross-validated kernel estimator").

If, instead of the kernel estimator, $f_{\lambda,n}$ is defined by

$$f_{\lambda,n}(x) = \frac{\lambda}{n} \sum_{j=-\infty}^{\infty} \chi_{[\frac{j-1}{\lambda},\frac{j}{\lambda})}(x)\{\sum_{i=1}^{n} \chi_{[\frac{j-1}{\lambda},\frac{j}{\lambda})}(x_i)\}$$

(a histogram with bin width $1/\lambda$, and an instance of the Method of Sieves), then exactly the same procedure defines a cross-validated smoothing parameter, and a resulting cross-validated estimator. $f_{\lambda,n}$ could in fact be any density estimator in which λ represents the degree of smoothing (possibly, λ is vector valued). For example, λ may be the parameter indexing the sieve, in which case cross-validation provides a fully data-defined Method of of Sieves estimator. And the setting is not limited to density

estimation; cross-validation applies to regression (with, for example, the "likelihood-like" expression (3) replaced by an analogous "sum of squared errors-like"expression) and many other estimation problems as well. Whatever the setting, cross-validation can be used to automatically smooth an estimator derived by the Method of Sieves.

3. Results about Cross-validated Estimator

We have studied the application of cross-validated estimators to both infinite and finite dimensional problems. In all cases, our motivation has been the usually good behavior of these estimators for small samples, as demonstrated by computer experiment (see subsection D below).

Our results on infinite dimensional estimation problems concern the consistency of cross-validated kernels and histograms for nonparametric density estimation. For details we refer to [ ] - but here, loosely stated, are the results: let $f_{\lambda,n}$ be either the histogram estimator with bin width $1/\lambda$ or the kernel estimator with a compact kernel and window width $1/\lambda$. Let λ_n^* be the corresponding cross-validated smoothing parameter, given n observations (see previous subsection for the definition of λ_n^*). If f, the target density, has compact support then $f_{\lambda_n^*,n}$ is strongly consistent in the L_1 metric, i.e.

$$\int_{-\infty}^{\infty} |f_{\lambda_n^*,n}(x) - f(x)|dx \to 0 \quad \text{a.s.}$$

Most instances of Method of Sieve estimators are only implicitly defined - as the solution of an optimization problem over a subset of the parameter space. The availability of an explicit representation for histograms and kernel estimators led us to believe that cross-validated versions of these estimators would be relatively easy to analyze. In fact, they were not. The proof of the above-stated consistency theorems is long and difficult. And it is probably true more generally that analysis of nonparametric estimators smoothed by data-driven techniques is extremely hard.

Some estimates for finite dimensional (parametric) problems also contain unspecified smoothing parameters, and these too can be data-defined by cross-validation. With the intention of using these more elementary examples to learn about cross-validation in general, and because finite dimensional applications of cross-validation are interesting in their own right, we have begun a mathematical study of such estimators. Consider, for example, the linear regression problem:

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i, \quad 1 \leq i \leq n \; \varepsilon_i \text{ iid } N(0,\sigma^2).$$

Or, in vector-matrix notation:

$$Y = X\beta + \varepsilon \quad \varepsilon \sim N(0,\sigma^2 I).$$

The least squares (maximum likelihood) estimator for β is

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

The ridge estimator for β is

$$\hat{\beta}_\lambda = (X^TX+n\lambda I)^{-1}X^TY \qquad \lambda \geq 0.$$

Observe that $\hat{\beta}_0$ is the least squares estimator. The introduction of λ into the least squares estimator may be motivated by any of the following considerations. (1) $\hat{\beta}_\lambda$ minimizes

$$\|Y-X\hat{\beta}\|^2 + n\lambda\|\beta\|^2.$$

Hence $\hat{\beta}_\lambda$ may be viewed as a penalized least squares estimator, with penalty for large values of $\hat{\beta}$. (2) When X^TX is "nearly singular" (poorly conditioned) $\hat{\beta}_0$ has large MSE due to the fact that the inverse of X^TX is involved in its derivation. Adding $n\lambda I$ to X^TX improves the conditioning and may be expected to reduce MSE. (3) Perhaps the best justification for ridge regression is the following easily demonstrated fact: for every n, β, and $\sigma^2 > 0$, there exists a $\lambda > 0$ such that

$$E\|\beta-\hat{\beta}_\lambda\|^2 < E\|\beta-\hat{\beta}_0\|^2.$$

Unfortunately, the optimal λ (in terms of MSE) depends on β and σ^2, so that we are again faced with a version of the smoothing problem.

Let us define a cross-validated version of the estimator $\hat{\beta}_\lambda$. Define $\hat{\beta}_\lambda^i$ to be the ridge estimator obtained by deleting the i'th observation. The squared error in predicting the i'th observation,

$$(y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_{\lambda j}^i)^2,$$

measures the appropriateness of λ as a smoothing parameter.

Define

$$L_\lambda = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} \hat{\beta}^i_{\lambda j})^2$$

and choose $\lambda = \lambda_n$ to minimize L_λ. The cross-validated ridge estimator (due to Allen [4]) is $\hat{\beta}_{\lambda_n}$. Our simulations (see subsection D below), and those of others (see, for example [30]) indicate that $\hat{\beta}_{\lambda_n}$ can be an extremely good estimator for β, especially when $X^T X$ is nearly singular or σ is large. In almost every case that we have studied, the mean squared error of the cross-validated ridge regression estimator is smaller than that of the ordinary least-squares estimator. Often, the ridge estimator reduces the MSE of least squares by 50 or more percent.

There is a closely related estimator, due to Golub, Heath, and Wahba [21], called the "generalized cross-validation" (GCV) ridge regressor. The GCV ridge regressor is computed by first rotating the coordinate system and then deriving the ordinary cross-validation (OCV) estimator. Simulations demonstrate that GCV generally performs somewhat better than OCV, and GCV is more easily computed and has proven to be more mathematically tractable. Although the analytic results mentioned below are for both GCV and OCV, we will not formally define the GCV estimator since this would require that we introduce somewhat involved notation.

Here, again loosely stated, is what we know about the analytic properties of cross-validated ridge estimators:[1] If $\hat{\beta}_{\lambda_n}$ is the GCV or OCV ridge regressor then

$$\|\hat{\beta}_{\lambda_n} - \beta\| \to 0 \quad \text{a.s.}$$

and $$(X^TX)^{1/2}(\hat{\beta}_{\lambda_n} - \beta) \sim N(0,\sigma^2 I).$$

Observe that for least squares the distribution of

$$(X^TX)^{1/2}(\hat{\beta}_0 - \beta)$$

is exactly $N(0,\sigma^2 I)$. Thus the cross-validated estimator asymptotically assumes all of the distribution properties of the ordinary regression estimator, while for small samples typically improving on ordinary regression (as measured by mean square error).

Principal component analysis is another modification of least squares regression that is used (among other reasons) for problems with ill-conditioned design matrices or small samples. It has been suggested (see [13], for a discussion in a slightly different context) that cross-validation may be an effective means for choosing the number of components retained in a principal component analysis. Our conclusions with regard to cross-validated principal component analysis parallel those for the cross-validated ridge regressor: simulations demonstrate its potential superiority over ordinary regression for finite samples,

---

[1] A more precise statement of this and other results for finite dimensional problems, together with proofs, will appear in a forthcoming manuscript by Geman and graduate student, Aytul Erdal.

and we can show that cross-validated principal component analysis is asymptotically equivalent to least squares. (For the latter, one simply shows that, almost surely, for all sample sizes sufficiently large the number of components chosen by cross-validation equals the real dimension of the regression surface.) Details will be in the manuscript by Erdal and Geman.

C. Studies on Specific Method of Sieves Estimators.

In addition to general questions of consistency for Method of Sieves estimators, we have made a study of individual properties for some particular instances of the method. Two examples are discussed.

1. Estimation of a Poisson Intensity Function.

One problem area that we have studied from several perspectives is the estimation of the intensity function of a nonhomogeneous two-dimensional Poisson process based on the observation of random projections of the points in a realization of the process. Motivation to study these problems comes from their direct application to computed tomography -- the reconstruction of the structure (intensity function) of a body in two or three dimensions on the basis of sets of projections of the body to one or two dimensions.

Our work to date has developed a basic understanding of theoretical properties of estimators derived from the method of sieves and the principle of maximum likelihood [33,34,35], convergence properties of computational algorithms for solving the very large scale optimization problem when maximum likelihood is used for

image reconstruction [1,34], and a software implementation for a microcomputer of two alternative algorithms for computing the m.l.e. [1].

The mathematical problem is related to single-photon emission tomography and to other nuclear medicine applications in section V below. Closely related applications and mathematical formulations are given in the recent work of Shepp and Vardi [45]. In our work, however, we have concentrated on a connection between the continuous spatial background Poisson process on one hand and estimates of its intensity function in the form of digital pictures, on the other hand. The relationship between the unknown function on a continuum and discretized estimates of it fits very naturally within the framework of the method of sieves.

The mathematical and inference problems take the following form. We describe the formulation in two dimensions; there are no essential changes in the general formulation or in the nature of the results in three dimensions, but the notation is a bit more cumbersome.

Let $\{(X_i,Y_i)\}_{i=1}^{N_T}$ be the N_T points of a realization over the time interval $[0,T]$ of a Poisson process with intensity function $\rho(x,y)$ per unit time. The function ρ is unknown and its estimation is our goal. The compact support Ω of ρ is known. For simplicity, we take Ω to be the unit square centered at $(0,0)$. The realization of the process is evolving with time t.

In fact, N_T is Poisson with mean $EN_T = T \iint_\Omega \rho(x,y)dxdy$. We assume ρ is in $L_1(\Omega)$.

The points (X_i,Y_i) are not directly observable. Rather, we observe a ray with an orientation Θ_i and distance S_i from the origin. The orientation Θ_i is uniformly distributed on $[0,2\pi)$; $\{\Theta_i\}_{i=1}^{N_T}$ are mutually independent and are independent of $\{(X_i,Y_i)\}$. The distance S_i is a function of X_i,Y_i and Θ_i; an oriented line L_i in direction Θ_i is passed through the point (X_i,Y_i) and S_i is the signed distance $S_i = -X_i \sin \Theta_i + Y_i \cos \Theta_i$ from the origin to L_i. Our observables are the N_T points $\{(\Theta_i,S_i)\}_{i=1}^{N_T}$. (In the single-photon tomography applications, one actually observes grouped data in place of the points (Θ_i,S_i), and effects from attenuation and photon scattering must be accounted for.)

The problem is then to estimate ρ on its domain Ω from observation of $\{(\Theta_i,S_i)\}_{i=1}^{N_T}$.

No a priori restrictions are imposed on ρ, except $\rho \in L_1(\Omega)$. Reasonable nonparametric estimators can be expressed through the method of sieves. One aspect of these formulations is distinctly different from problems analyzed previously: the observables, even when conditioned on N_T, do not constitute an iid sample with intensity (viz. density) ρ. Instead, the observables represent an iid sample from a density proportional to $R\rho$, where R is an integral operator on $L_1(\Omega)$. To estimate ρ, we need to address the problems of "inversion" of R, topological and operator theoretic problems such as identifying the appropriate

domain for R, demonstrating that it is invertible on the image of that domain, and understanding continuity properties of R and R^{-1}.

Results about consistency of nonparametric estimators of ρ and about characterization of algorithms for computing estimators in special cases are reported in [34,35]. First we confirm that $\{(\Theta_i,S_i)\}_{i=1}^{N_T}$ constitutes a realization of a Poisson process on a region $\hat{\Omega}$; the intensity function of the process is

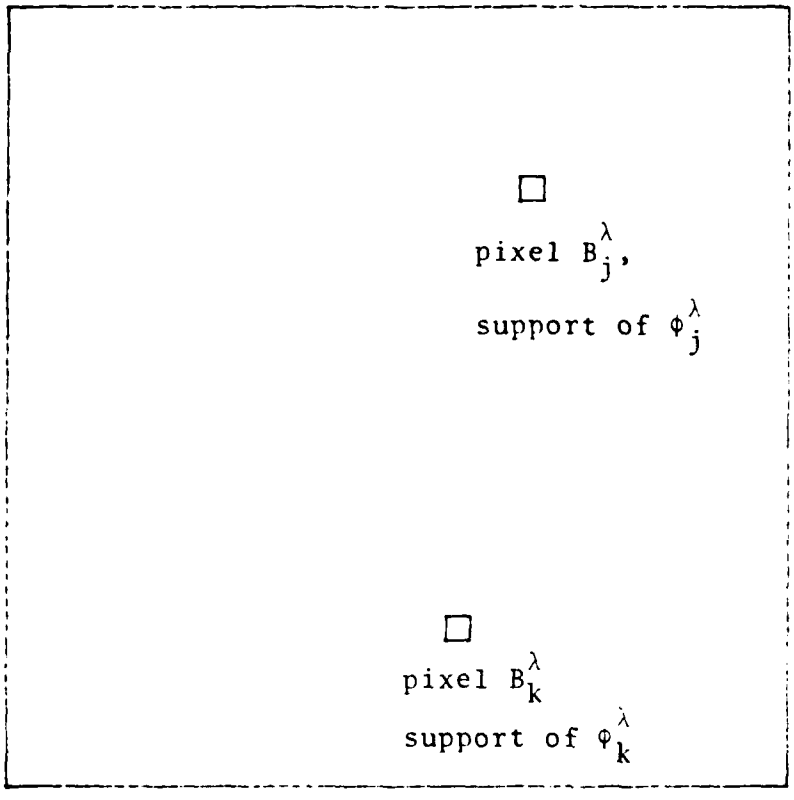
$$\hat{\rho}(\theta,s) = \frac{1}{2\pi} T \int_{L_\theta(s)} \rho(x,y)d\ell = \frac{T}{2\pi} R\rho(\theta,s),$$

where the line integral is taken over the straight line path $L_\theta(s)$ passing through (x,y) having positive orientation θ, and satisfying $s=-x \sin\theta+y \cos\theta$; ℓ denotes arc length on $L_\theta(s)$. The transform domain Ω is depicted in the Figure 1. It is the set of all (θ,s) for which the corresponding line intersects the square Ω. R is the Radon transform [27].

In the general approach followed in [35], the main steps are (i) to specify a sieve and estimate $\hat{\rho}$ by maximizing the likelihood of $\{(\Theta_i,S_i)\}$ over the sieve, and (ii) to map the m.l.e. $\hat{\rho}^*$ of $\hat{\rho}$ into an estimator ρ^* of $\rho = R^{-1}\hat{\rho}$.

The first step is reasonably straightforward, except for purely technical problems associated with the domain $\hat{\Omega}$ and the unusual form relative to $\hat{\Omega}$ of natural sieves. The likelihood function assumes a convenient form for analysis:

$$\mathcal{L}(\hat{\rho}) = \frac{\mathcal{J}^{N_T}}{N_T!} e^{-\mathcal{J}} \prod_{i=1}^{N_T} \left(\frac{\hat{\rho}(\Theta_i,S_i)}{\mathcal{J}}\right),$$



domain Ω =
$[-.5,.5]\times[-.5,.5]$

transform domain
$\hat{\Omega}$

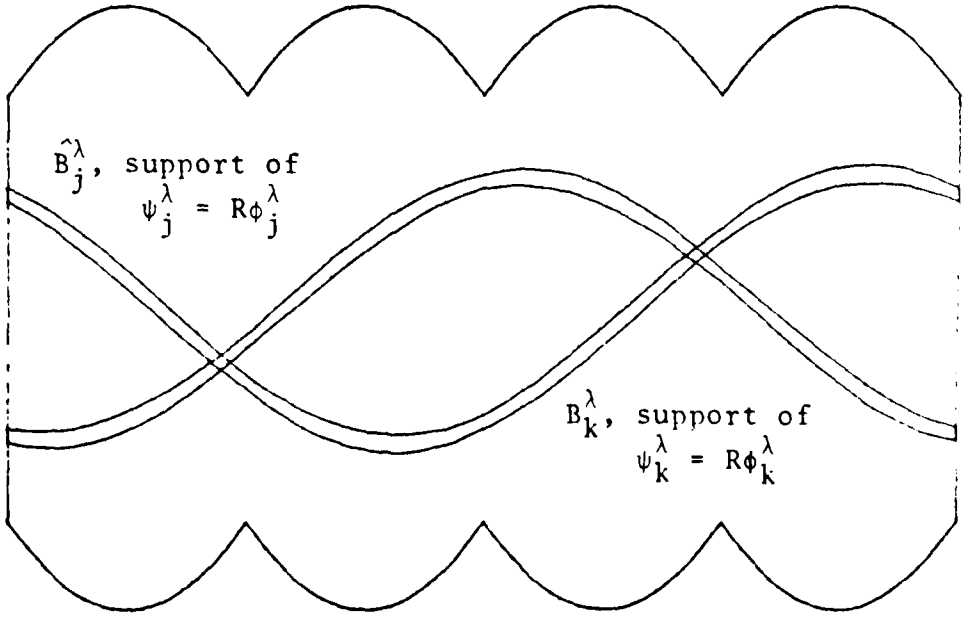


Figure 1

where $\mathcal{I} = \iint_{\hat{\Omega}} \hat{\rho}(\theta,s)d\theta ds$.

The most "natural" sieve, thinking in terms of potential applications to emission tomography and construction of digital pictures, is based on a partition of Ω into λ congruent square pixels. Let B_j^λ denote the sets in this partition and let ϕ_j^λ be the indicator function of B_j^λ, $j=1$ to λ. Then

$$S_\lambda = \{\rho \in L_1(\Omega): \quad \rho = \sum_{j=1}^{\lambda} a_j \phi_j^\lambda\}.$$

The image of S_λ in $L_1(\hat{\Omega})$ determines the sieve

$$\hat{S}_\lambda = \{\hat{\rho} \in L_1(\hat{\Omega}): \quad \hat{\rho} = \sum_{j=1}^{\lambda} a_j \psi_j^\lambda\},$$

where $\psi_j^\lambda = R\phi_j^\lambda$. Each of the basis functions $\psi_j^\lambda(\theta,s)$ for $\hat{S}_\lambda$ has a support inside $\hat{\Omega}$ in the shape of a narrow sinusoidal band (see Figure 1); the support of ψ_j^λ consists of the oriented lines (θ,s) that intersect the pixel B_j^λ. Within its support, for any fixed θ_0, $\psi_j^\lambda(\theta_0,s)$ is a piecewise linear function of s.

Now $\hat{\rho}$ is estimated by fixing $\lambda = \lambda_T$ and finding a maximizing function $\hat{\rho}_T^*$ in $\hat{S}_\lambda$ of the likelihood $\mathcal{L}(\hat{\rho})$. Except for the unusual structure of the basis functions ψ_j^λ, this is a straightforward implementation of the method of sieves. We can show, using the "small-ball technique" on $L_1(\hat{\Omega})$ that if

(i) $\lambda_T = o(T^{1/4-\varepsilon})$ for some $\varepsilon > 0$, and

(ii) $\iint_{\hat{\Omega}} \hat{\rho}(\theta,s) \ell n\, \hat{\rho}(\theta,s)d\theta ds < \infty$, then

$\lim_{T\to\infty} \iint_{\hat{\Omega}} |\hat{\rho}(\theta,s) - \hat{\rho}_T^*(\theta,s)|d\theta ds = 0$ with probability one.

Condition (i) is, we feel, not sharp, and we are continuing to investigate this.

It is a delicate matter to translate the consistency of $\hat{\rho}_T^*$ into a consistency theorem for $R^{-1}\hat{\rho}_T^* = \rho_T^*$. Indeed, R^{-1} is not a continuous mapping from $L_1(\hat{\Omega})$ into $L_1(\Omega)$. We can, however, regard R as an operator on a domain other than $L_1(\Omega)$ and obtain useful continuity properties. Two approaches are possible, based on results about the Radon transform developed by Ludwig [31]: (i) we can restrict R to a domain smaller than L_1, impose strong regularity conditions and a stronger topology on that domain and obtain a useful continuity property for R, or (ii) we can extend R to a domain larger than L_1, relax conditions on elements of the domain, prescribe a weaker topology for the enlarged domain, and we obtain a bicontinuous extended operator R. The first approach requires us to impose unrealistic smoothness restrictions on the intensity function ρ being estimated, so the usefulness of theorems derived by this approach is limited. The second approach adheres to the spirit of truly nonparametric inference by not imposing additional restrictions on ρ. The price paid by the theory is that consistency results are expressed in a weaker topology, that is, we will conclude $\rho_T^* \to \rho$ in a weak topology as $T \to \infty$ (but still in a strong sense probabilistically, i.e. almost surely). The approach bears a strong resemblance to methods used to prove existence of solutions of PDEs by first confirming existence of a solution in a weak sense and then analyzing the regularity properties of the weak solution.

For the present problem, starting with $L_1(\Omega)$ as our space for ρ, it is natural to use an extension of R to the space $\mathcal{S}'$ of Schwartzian distributions with compact support. Any element of $L_1(\Omega)$ is such a generalized function since $\iint_{R^2} \rho(x,y)\phi(x,y)dxdy$ is a continuous linear functional on the space $\mathcal{S} = C^\infty(\mathbb{R}^2)$ of test-functions ϕ. Details of the extension of R are given in [27] and [31]. The extended R is a bicontinuous bijection from $\mathcal{S}'$ to a completely characterized subspace $\mathcal{S}' \cap N^\perp$ of $\mathcal{S}'$.

Now we can translate convergence of $\hat{\rho}_T^*$ into convergence of ρ_T^*. The strong norm convergence of $\hat{\rho}_T^*$ to $\hat{\rho}$ implies convergence in the weaker topology of the range $\mathcal{S}' \cap N^\perp$, which in turn, from continuity of R^{-1}, implies convergence of ρ_T^* to ρ. Specifically, for any ϕ in $C^\infty(\mathbb{R}^2)$, $\lim_{T\to\infty} \iint_{\mathbb{R}^2} \rho_T^*(x,y)\phi(x,y)dxdy = \iint_{\mathbb{R}^2} \rho(x,y)\phi(x,y)dxdy$ with probability one.

This general analytical strategy is applied in [35] to determine consistency of estimators based on alternative sieves.

Other results are reported in [34] where we consider intensity functions ρ which are piecewise constant on subsets of Ω and we wish to estimate the unknown sets. We relate the set estimation problems to more familiar problems of estimating a univariate unimodal density function. That connection allows us to adapt consistency and asymptotic distribution results [38, 57], to characterize nonparametric m.l.e.s in terms of

convex envelopes of counting functions of the projected point process, and to obtain direct computational algorithms for the m.l.e. along with results on the complexity of the algorithms.

In directly related work, graduate student Nicholas Accomando has designed and implemented algorithms on a microcomputer for computation of estimators of ρ based on maximum likelihood and the method of sieves. The algorithms are designed with the single-photon emission tomography problems in mind and they allow for effects of photon attenuation. Thus they implement inversion of the more general attenuated Radon transform R_μ (see section V) of which R is a special case. The software developed for the Data General MP-200 microcomputer uses a combination of PASCAL and assembler code. The assembler language modules are documented in PASCAL in order to facilitate implementation on other computers. The programs implement both gradient methods of calculating the m.l.e. and the EM algorithm analyzed and used by Shepp and Vardi [45]. The software design problem is a hard one since it requires the balancing of the constraints of a small computer and of a very big optimization problem. Accomando's system can handle the method-of-sieves estimator with $\lambda = 8100$ free parameters, corresponding to a 90×90 digital picture; the MP-200 has 32k 16-bit words of high speed memory to which Accomando has added a board with 64k 16-bit words. The programs will reconstruct a 60×60 phantom in about fifteen or twenty minutes, depending on the algorithm (EM or conjugate-gradient) and on the number of iterations needed.

Accomando's implementation is convincing proof of the feasibility of maximum-likelihood and the method-of-sieves for solving large scale problems with readily available computational resources.

2. Convergence Rates and Asymptotic Distribution for some Canonical Sieves.

The general question of convergence rates and asymptotic distribution for sieve estimators is extremely difficult. Professor H.T. Nguyen, in personal correspondence, has suggested that estimators based on sieves which consist of increasing subspaces of a Hilbert space would be particularly amenable to thorough mathematical analysis. A recent paper by Nguyen and Pham [37] puts this idea to good use. They utilize a sieve of this type to estimate the drift frunction of a repeatedly observed non-stationary diffusion, and they are able to partially specify rates of convergence and asymptotic distribution.

Here is a more elementary example of Nguyen's idea, which we have explored in some detail. Recall the nonparametric regression problem discussed briefly in section II. Let us here take x, the "independent" variable, to be deterministic. We then think of the distribution of Y as being an unknown function of x, $F_x(\cdot)$. For this example, we will assume $x \in [0,1]$. The problem is then to estimate

$$\alpha_0(x) = E_x[Y] \equiv \int_{-\infty}^{\infty} y F_x(dy) \qquad x \in [0,1]$$

from independent observations $y_1,\ldots y_n$, where $y_i \sim F_{x_i}$, and $x_1,\ldots x_n$ is a deterministic, so-called design, sequence. In other words, for each i=1,2,...n we make an observation, y_i, from the distribution F_{x_i}, and from these observations we wish to estimate the mean of Y as a function of x. For a specific example, let us assume that the design sequence, for fixed n, is equally spaced on the interval [0,1]:

$$x_j = \frac{j}{n} \quad j=1,2,\ldots n.$$

Of course, an unconstrained minimization of the sum of square errors

$$\sum_{i=1}^{n} (y_i - \alpha(x_i))^2$$

will not produce a useful estimator. Using the Hilbert space $L_2([0,1],dx)$, a sieve of the type suggested by Nguyen is the "Fourier sieve"

$$S_m = \{\alpha(x):\alpha(x) = \sum_{k=-m}^{m} a_k e^{2\pi ikx}\};$$

it is particularly tractable and makes for a good illustration of his idea. The sieve size is governed by the parameter m, which will be allowed to increase to infinity with n. If we restrict m_n so that $m_n \leq n$ for all n, then $\hat{\alpha}_n$ is uniquely defined by requiring that it

$$\text{minimize } \sum_{i=1}^{n} (y_i - \alpha(x_i))^2 \text{ subject to } \alpha \in S_{m_n}.$$

Because of the subspace character of the sieve, it is quite easy to show that for any sequence $m_n \uparrow \infty$ such that $m_n/n \to 0$ and $m_n \leq n$,

$$E \int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2 dx = 0(\frac{1}{m_n} + \frac{m_n}{n} + \frac{1}{\sqrt{n}})$$

as $n \to \infty$. In particular, if $m_n \sim \sqrt{n}$, then

$$E \int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2 dx = 0(\frac{1}{\sqrt{n}})$$

as $n \to \infty$. (All under some very mild assumptions - see [18] for details.)

What does the least squares estimator, $\hat{\alpha}_n$, look like? A simple calculation gives the explicit form:

$$\hat{\alpha}_n(x) = \frac{1}{n} \sum_{i=1}^{n} y_i D_{m_n}(x-x_i)$$

where D_m is the Dirichlet kernel

$$D_m(x) = \sum_{|k| \leq m} e^{-2\pi ikx} = \frac{\sin \pi(2m+1)x}{\sin \pi x} .$$

Here, then, the least squares (sieve) estimator turns out to be a kernel estimator. Kernel estimators for nonparametric regression have been widely studied, although from a somewhat different point of view. See [3], [11], [16], [43] and [48] for some recent examples. It is not too difficult to now exploit this simple form for $\hat{\alpha}_n$, and say a good deal more about its behavior. Let

$$V(x) = \int_{-\infty}^{\infty} (y-\alpha_0(x))^2 F_x(dy),$$

the variance of Y at x. Then if $\alpha_0(0) = \alpha_0(1)$, and if $m_n \uparrow \infty$, such that $m_n = 0(n^{\beta})$ for some $\frac{1}{4} < \beta < \frac{1}{2}$, the process

$$p_n(t) \equiv \sqrt{n} \int_0^t (\hat{\alpha}_n(x) - \alpha_0(x))dx$$

converges weakly on [0,1] to the diffusion, p(t), defined by

$$dp(t) = \sqrt{V(t)}\, dW_t, \quad p(0) = 0$$

where W_t is standard Brownian motion. (Again, see [18] for details.)

The conditon $\alpha_0(0) = \alpha_0(1)$ is awkward, but it cannot be removed. It is a consequence of the sieve, S_n, which admits only functions which are continuous on the unit torus. Another sieve of the subspace type, more natural in the absence of the assumption $\alpha_0(0) = \alpha_0(1)$, is

$$S_n' = \{\alpha(x): \alpha(x) = \sum_{k=-m}^{m} a_k \cos[k \arccos(2x-1)]\}$$

i.e. replace the trigonometric polynomials by the Chebyshev polynomials. Here we would want to choose a design sequence which preserves the orthogonality of the basis sequence:

$$x_j = \frac{1}{2} + \frac{1}{2} \cos[2j-1)\pi/2n], \quad j=1,2,\ldots n.$$

We presume that the above results have their analogues for S_n' as well.

Still a good deal more can be said about the estimator $\hat{\alpha}_n$.

With suitable restrictions on the growth of m_n, we can establish: pointwise convergence ($E|\hat{\alpha}_n(x)-\alpha_0(x)|^2 \to 0$ for each $x\in(0,1)$); pointwise asymptotic normality; and a relation between the smoothness of α_0 and the rate at which

$$E\int_0^1 |\hat{\alpha}_n(x)-\alpha_0(x)|^2 dx \text{ converges to zero.}$$

D. Computer Experiments

One of the most powerful strategies for our past research on the method of sieves has been to systematically use computer experimentation. The experiments often involve simulation of the processes and inference methods which we are analyzing. And frequently the experiments do not involve simulation per se, but use the machine to solve numerically the analytical hurdles that come up in a purely theoretical analysis. Both kinds of experiments have been invaluable complements to our analysis. They have provided direction and reinforced confidence in the day-to-day evolution of the theoretical work; they have made it possible for us to understand and to demonstrate connections between asymptotic theorems on one hand and small-sample properties of method-of-sieves estimators on the other hand; and they have frequently suggested hypotheses to us, providing new ideas to be confirmed later by rigorous analysis.

Most of our experiments have been carried out interactively, using a substantial library of APL programs developed by Grenander and McClure [25] as a tool for experimental mathematics.

The library has been assembled systematically over the past four years. It is well documented and is now being shared with mathematicians elsewhere. The examples of experiments described below have been facilitated by the availability of the programs for interactive graphics, simulation, linear algebra and matrix spectrum analysis, computational geometry, quadrature, and discrete Fourier analysis.

Example 1. The theoretical understanding of cross-validated ridge regression is essentially complete, at least the asymptotic properties are clear, since the cross-validated estimator $\hat{\beta}_{\lambda_n}$ has been shown to have the same asymptotic distribution as the least-squares estimator of β. Extensive simulations have been useful for demonstrating (i) the small-sample behavior of $\hat{\beta}_{\lambda_n}$, (ii) the relationship between ordinary and generalized cross-validation, (iii) the sensitivity of the performance of $\hat{\beta}_{\lambda_n}$ to ill-conditioning of X^TX, and (iv) the dependence for small-samples of the mean-squared-error of $\hat{\beta}_{\lambda_n}$ on the error variance σ^2.

Two examples of the simulation results are depicted in Figures 2 and 3. The ordinate on each graph is a so-called Relative Error for a cross-validated (ordinary or generalized) ridge regression estimator $\hat{\beta}_\lambda$ of β in the model $Y = \beta^TX+\varepsilon$, where X is $n\times p$ with $n=30$ and $p=5$, the variance of ε is $\sigma^2 I$ ($\sigma=.01$ is Figure 2), and α is an index of the ill-conditioning of the 5×5 matrix X^TX ($\alpha=0.8$ in Figure 3); the eigenvalues of X^TX are, on the average, $1-\alpha^2$ with multiplicity four and $1+4\alpha^2$

Relative Errors for Cross-Valdiated Estimators $\hat{\beta}_\lambda$.

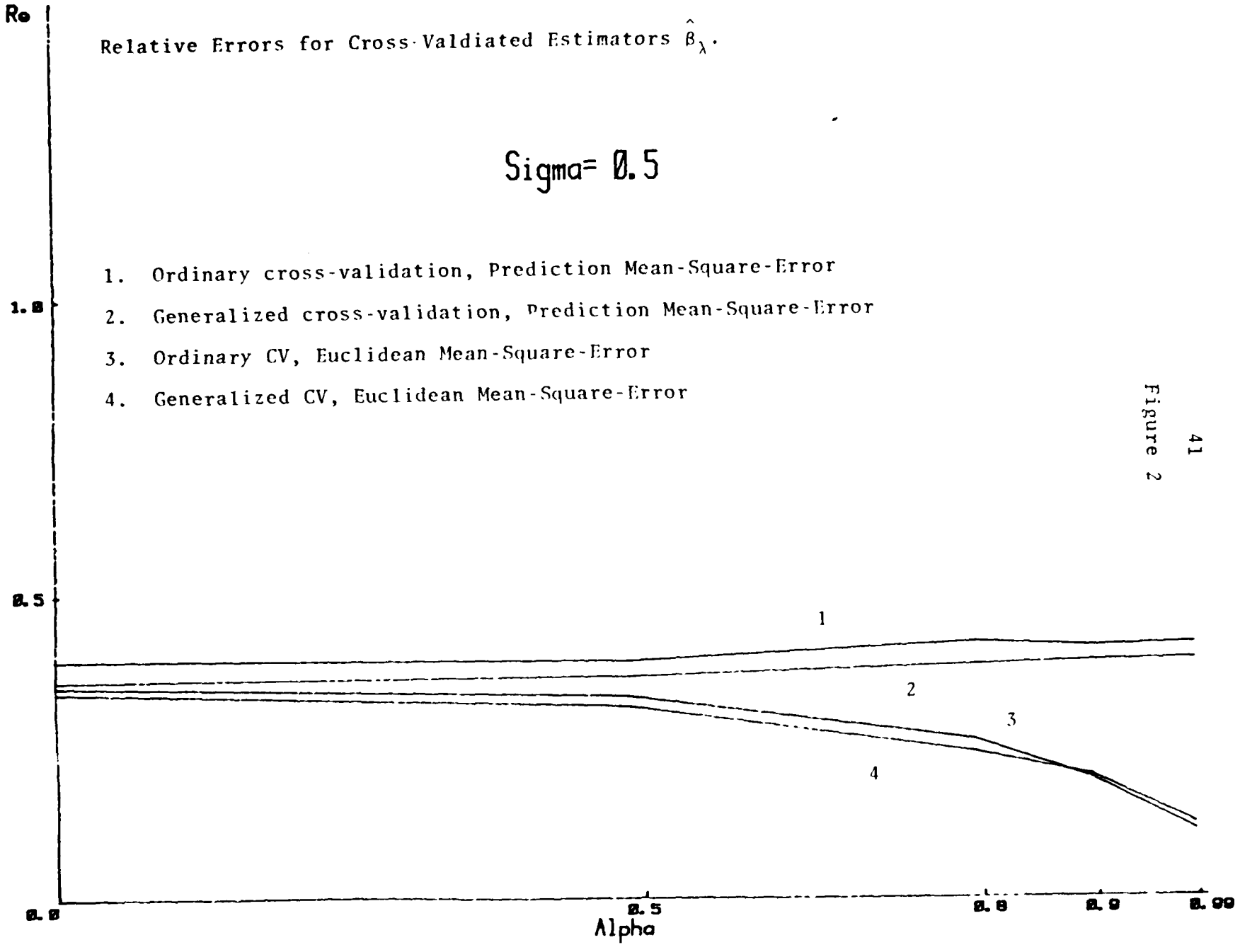


Figure 2

Relative Errors for Cross-Validated Estimators $\hat{\beta}_\lambda$.

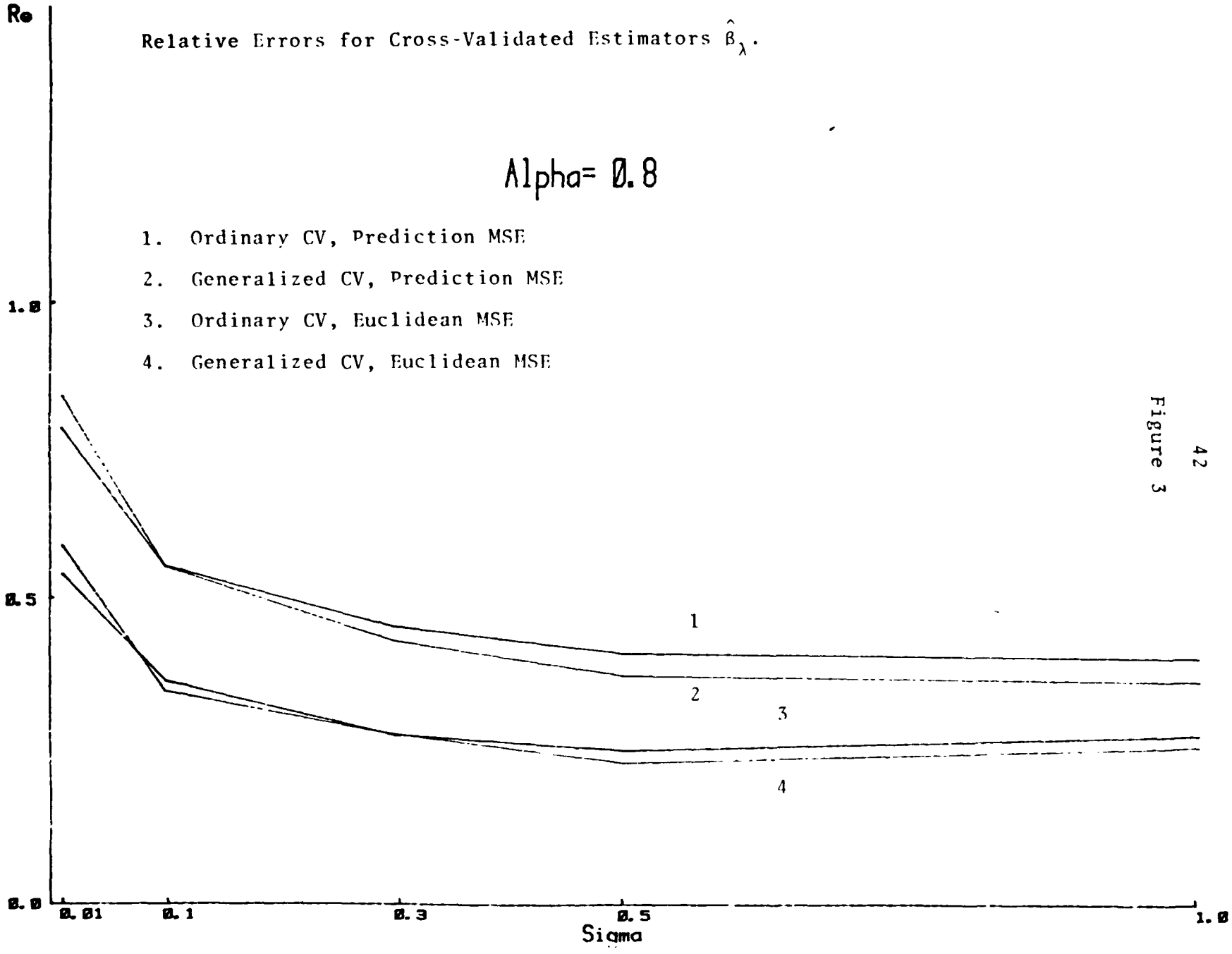


Figure 3

with multiplicity one, so X^TX collapses to a rank one matrix as α goes to 1.

The precise descriptions of the two mean-squared-errors, prediction and Euclidean, used to assess $\hat{\beta}_\lambda$ are not particularly important for describing the general conclusions that can be drawn from these experiments. It suffices to know that the Relative Errors are ratios of a mean-squared-error for the cross-validated estimator to that mean-squared-error for the least squares estimator of β. The asymptotic efficiency of the least-squares estimate when ε is Gaussian together with the asymptotic distribution theorem for $\hat{\beta}_\lambda$ (see p.25) combine to show that for any fixed values of the parameters α and σ, the Relative Error goes to 1 as n goes to ∞. The difference from 1 for these small sample (n=30) experiments is striking. Both ordinary and generalized CV estimators perform significantly better than the asymptotically efficient least-squares estimator. The results also exhibit the greater advantage of ridge regression with cross-validation as X^TX becomes more singular.

These simulations were carried out fairly early during Erdal and Geman's study of properties of cross-validated ridge regressors and the experiments had a direct influence on the direction of the theoretical developments. The pictures provided convincing evidence that the ridge regressors are in general as good as (and usually better than) ordinary least squares. In addition, the relative small differences in simulation results for the ordinary compared to the generalized cross-validated estimators stimulated the successful and more intricate analysis of the ordinary cross-validated estimator.

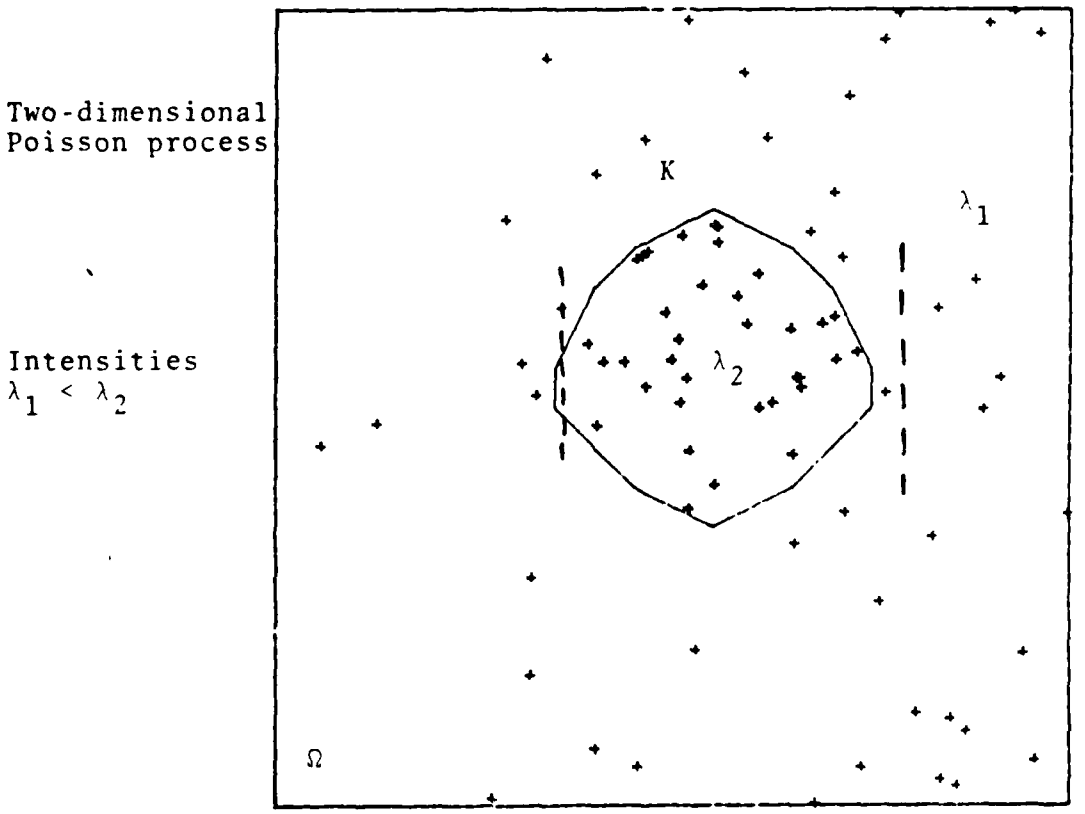
Example 2. Experiments have been carried out in connection with method-of-sieves estimates of two-dimensional Poisson intensity functions and B-spline estimates of univariate density functions. The two estimation problems are directly related to each other. The experimentshave been used to (i) determine the small sample behavior of estimators whose consistency is reasonably straightforward to prove and (ii) explore questions of consistency for cross-validated estimators, whose consistency properties are not yet known.

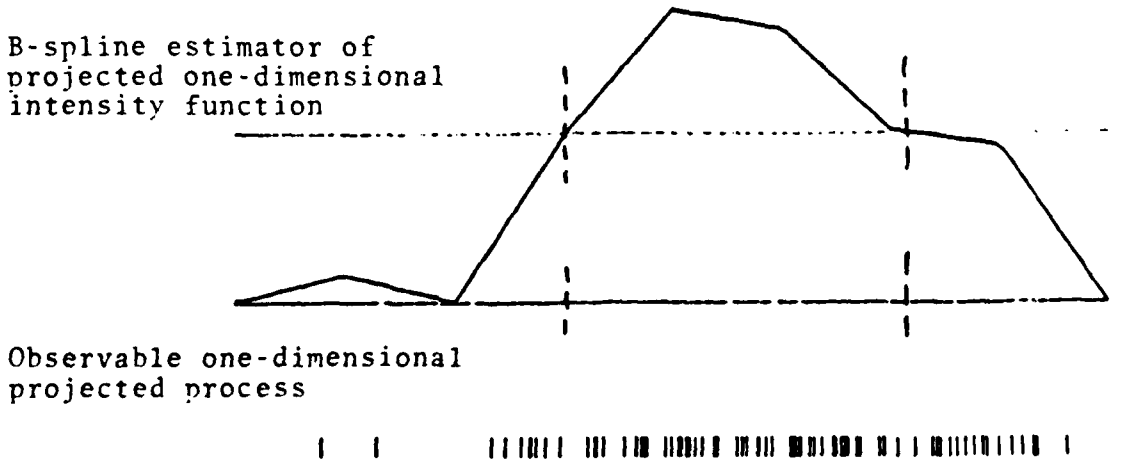
Figure 4 depicts a realization of a two-dimensional Poisson process on the square Ω. For this particular experiment, the nonuniform intensity function assumes two values, corresponding to a "hot" region $K \subseteq \Omega$ surrounded by lower-level background Poisson events. The unknown region K is assumed to be convex, but otherwise arbitrary and the goal is to reconstruct K from a finite set of independent projections of the Poisson points.

The points plotted at the bottom of the figure are projections to the x-coordinates alone of the Poisson points in Ω. The projected points are a realization of a one-dimensional Poisson process with a nonhomogeneous intensity that is unimodal and nonuniform in the interval "shadow" of K. An estimation procedure analyzed in [34] is based on using the univariate process to infer the location of K's shadow. In the particular case of polygonal K, the univariate intensity function is a spline of order 2 (piecewise linear) and spline estimators are natural.

id: 4-81, s02
sample size is 81
lam2 : lam1 = 10 : 2

Figure 4





The function depicted in Figure 4 is an estimate for this particular realization of the process of the univariate intensity. It is a linear combination of B-splines with equally-spaced knots. Once the number of knots is specified, the coefficients of the basis functions that determine the estimate are found by maximizing the likelihood function. The general recipe for establishing consistency [19] applies here to assure that the B-spline estimators are consistent if the number of knots goes to ∞ suitably slowly ($0(n^{1/4-})$) as sample size n increases. But for a fixed-size sample, the problem of specifying the number of spline knots is another instance of the recurrent smoothing problem in nonparametric estimation.

The spline estimate in Figure 4 lets the data prescribe the appropriate number of knots by half-sample cross-validation, maximizing a cross-validation likelihood with respect to the number of knots; <u>one-half</u> the sample is deleted for each of <u>two</u> factors in the cross-validation likelihood, instead of deleting one sample unit for each of n factors as in eqn. III.B.2.3.

The experiments that we have done suggest that the half-sample cross-validated B-spline estimators are effective for problems such as this one where a target density (or intensity) has compact support. The theoretical confirmation of any consistency properties remains a tempting and elusive problem. The gap between this problem and known results occurs because the estimators here are not prescribed explicitly, but only implicitly through the principle of maximum likelihood.

The experimental results present compelling evidence that consistency results analogous to known ones will hold for much more general types of estimators than ones that admit explicit representation. Another instance is depicted in Figure 5. Here a cubic spline with m equally spaced knots is used to estimate a density function for a data set used by Boneva, Kendall and Stefanov [7] to illustrate alternative spline estimators. The B-spline coefficients are determined by maximum likelihood and the number of knots is found by maximizing the half-sample cross-validation likelihood. Our experience with the experiments encourages continuing analysis of the theoretical problems as described in section IV.A.2 below.

Experiments that are similar in spirit to the two described above have been instrumental in focusing our attention on important aspects of the theoretical analysis of (i) the consistency problem for cross-validated histograms and kernel estimators [10] and (ii) complex operator equations that characterize the bias of certain estimators of regression functions of two variables (see IV.B). For the cross-validation problem, the experiments reported by Schuster and Gregory [42] pointed to the pivotal role of spacings in deciding the consistency of density estimators. Our own simulations are important in the formulation of simple sufficient conditions for consistency. For the surface regression problems, the striking simplicity of the numerical solutions of very large systems of linear equations pointed the direction for proving general properties of those solutions.

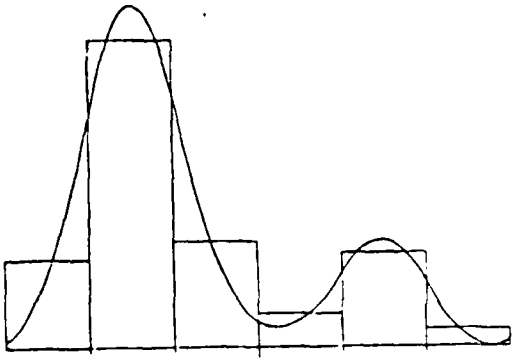
ID: STEPHENS

Figure 5

N = 76

K = 4 ; L = 12

CVLL = -418



BKS Histospline Estimator
From [7]

FIG. 7

Cross-Validated Cubic B-spline Estimator

Stephen's Data on Orientation of 76 Turtles.

IV. Projects in Progress

As of June 1, 1983 a number of projects were in progress, at varying stages of completion, that relate directly to our research program on nonparametric estimation by the method of sieves. These projects involve collaboration of the principal investigators with other faculty members (Ulf Grenander at Brown, Lewis Pakula at the University of Rhode Island, and Donald Geman at the University of Massachusetts), with postdoctoral associates (Barry Davis and Aytul Erdal), and with Ph.D. candidates (Shoulamit Shwartz, Nicholas Accomando, Joyce Anderson, Edmond Nadler and Brock Osborn). We shall briefly mention the status of these projects, since the work to date has been supported by contract DAAG29-80-K-0006.

1. Aytul Erdal and Stuart Geman have prepared drafts of manuscripts that report new consistency and asymptotic distribution results for cross-validated estimators in ridge regression and principal components analysis. The manuscripts are based on results reported in Erdal's 1983 Ph.D. dissertation. The papers will be submitted for publication.

2. Stuart Geman, Donald McClure, and Barry Davis continue active collaboration with members of the Division of Cardiology at Rhode Island Hospital on several digital image processing problems. These include image registration methods for noninvasive digital subtraction imaging of coronary arteries and image enhancement and restoration methods for nuclear medicine scans of the myocardium that have been degraded by photon attenuation. We are experimenting with the use of the method of sieves to estimate nonlinear

transformations that map one scene into another; we believe that this will prove to be an effective and versatile method for registration of scenes.

Joyce Anderson is working with McClure on use of maximum likelihood and the method of sieves to enhance the attenuated nuclear medicine scans. These applications have motivated the study of new theoretical problems in nonparametric estimation based on censored data and incomplete observables.

3. Shoulamit Shwartz has completed the research for her Ph.D. thesis on the design of triangulation sieves for nonparametric multiple regression. The theoretical questions are motivated by surface reconstruction problems in remote sensing. The thesis is in draft form and the results will be included in manuscripts being prepared for publication with Donald McClure.

4. Stuart Geman and Donald McClure are working together with Ulf Grenander, Lewis Pakula and Donald Geman on a project to develop mathematical models for complex systems. The phenomena being modeled include real pictures of outdoor scenes and of highly structured scenes and systems as disparate from digital images as neural systems for complex decision processes such as medical diagnosis. To date, nine internal working papers have been prepared by the group. These include a master plan for the project, an outline of specific open problems, and preliminary results on particular complex systems under investigation.

5. Stuart Geman, Donald McClure and David Cooper recently organized the program for an ARO sponsored Workshop on Unsupervised Image

Analysis. The workshop was held at Brown on 14-16 April 1983. McClure and Cooper are currently working with Dr. Robert Launer on the editing of the workshop proceedings which will be published by Academic Press. Included in the proceedings will be papers by Geman on "A Markov random field model for image segmentation" and by McClure on "Image processing algorithms based on methods of nonparametric inference".

6. Stuart Geman presented an invited talk at the June meeting of the Institute of Mathematical Statistics in Arcata, CA on the work with Grenander, D. Geman, and McClure on hierarchical Markovian models for discrete pictures and restoration algorithms based on these models.

7. Stuart Geman will present an invited talk on "A parallel realization for maximum entropy distributions with applications to problems in inference and optimization" at the Third Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics on 1-4 August 1983 at the University of Wyoming.

8. Donald McClure will present an invited talk on mathematical experiments on the computer at the Institute of Mathematical Statistics Special Topics Meeting on Statistics and Computing on 24-26 October 1983 at Pennsylvania State University.

V. Publications and Technical Reports

The list in this section includes internal and interim technical reports and published papers that have been prepared as part of the project on nonparametric estimation by the method of sieves. Technical reports in the series "Reports in Pattern Analysis" can be obtained from the Division of Applied Mathematics at Brown University.

1. S. Geman & C-R. Hwang, Nonparametric maximum likelihood estimation by the method of sieves, Reports in Pattern Analysis No. 80, 1979. Ann. Statist. vol. 10, 1982, 401-414.

2. S. Geman & C-R. Hwang, A chaos hypothesis for some large systems of random equations, Reports in Pattern Analysis No. 82, 1979. Z. Wahrscheinlichkeitstheorie verw. Gebiete, vol. 60, 1982, 291-314.

3. S. Geman, An application of the method of sieves: functional estimator for the drift of a diffusion, Reports in Pattern Analysis No. 92, 1980. Colloquia Mathematica Societas Janos Bolyai, vol. 32, 1980, 231-252.

4. C. Plumeri, Conditioning by inclusion when connection type is LINEAR, Reports in Pattern Analysis No. 94, 1980. (Ph.D. work supervised by D. McClure)

5. S. Geman, The law of large numbers in neural modelling, Reports in Pattern Analysis No. 95, 1980. SIAM-AMS Proceedings, vol. 13, 1981, 91-105.

6. C. Plumeri, On convergence of sums of Markov random variables, Reports in Pattern Analysis No. 96, 1980. (Ph.D. work supervised by D. McClure)

7. S. Geman, Almost sure stable oscillations in a large system of randomly coupled equations, Reports in Pattern Analysis No. 97, 1980.

8. S. Geman, Sieves for nonparametric estimation of densities and regressions, Reports in Pattern Analysis No. 99, 1981.

9. C. Plumeri, Asymptotic probability measure on regular structures, Reports in Pattern Analysis No. 100, 1981. (Ph.D. dissertation supervised by D. McClure)

10. J. E. Anderson, Experiments with the method of sieves for density estimation, Reports in Pattern Analysis No. 105, 1981. (Honors thesis supervised by S. Geman & D. McClure)

11. Y-S. Chow, S. Geman & L-D. Wu, Consistent cross-validated density estimation, Reports in Pattern Analysis No. 110, 1981. Ann. Statist. vol. 11, 1983, 25-38.

12. S. Geman & D. E. McClure, Characterization of a maximum likelihood nonparametric density estimator of kernel type, Reports in Pattern Analysis No. 114, 1982. Proceedings of the NASA Workshop on Density Estimation and Function Smoothing, L. F. Guseman, Jr. (editor), Texas A&M University (1982), 38-47.

13. D. E. McClure, Estimation of planar sets from Poisson projections, Reports in Pattern Analysis No. 115, 1982. Proceedings

of the NASA Workshop on Density Estimation and Function Smoothing, L. F. Guseman, Jr. (editor), Texas A&M University (1982), 38-47.

14. W. B. Levy & S. Geman, Limit behavior of experimentally derived synaptic modification rules, Reports in Pattern Analysis No. 121, 1982.

15. S. Geman, Method of sieves, Reports in Pattern Analysis No. 125, 1982. Encyclopedia of Statistical Sciences, vol. 5, 1983.

16. N. Accomando, The implementation of maximum likelihood reconstruction algorithms for single photon emission tomography, Reports in Pattern Analysis No. 126, 1982. (Ph.D. work supervised by D. McClure)

17. A. Erdal, Cross-validated ridge regression, Reports in Pattern Analysis No. 127, 1982. (Ph.D. work supervised by S. Geman)

18. A. Erdal, The method of cross-validation for principal component analysis, Reports in Pattern Analysis No. 128, 1983. (Ph.D. work supervised by S. Geman)

19. B. R. Davis & S. Geman, The application of neurobiological and statistical concepts to machine intelligence, Reports in Pattern Analysis No. 129, 1983.

20. E. Nadler, Least square approximation of functions on an equilateral triangle by linear functions, Reports in Pattern Analysis No. 131, 1983. (Ph.D. work supervised by D. McClure)

21. S. Shwartz, Optimal design of triangulation sieves for

nonparametric multiple regression and surface restoration, in preparation. (Ph.D. dissertation supervised by D. McClure)

22. S. Geman, A Markov random field model for image segmentation, in preparation, to appear in Automated Image Analysis: Theory and Experiments, D. Cooper, R. Launer & D. McClure (editors), Academic Press.

23. D. E. McClure, Image reconstruction algorithms based on methods of nonparametric inference, in preparation, to appear in Automated Image Analysis: Theory and Experiments, D. Cooper, R. Launer & D. McClure (editors), Academic Press.

VI. Personnel

The following people have made substantive contributions to the research project on nonparametric estimation supported by contract DAAG29-80-K-0006.

Stuart Geman, co-principal investigator.

Donald E. McClure, co-principal investigator.

Chii-Ruey Hwang, collaborator with S. Geman and Visiting Assistant Professor at Brown University, 1981.

Nicholas Accomando, Ph.D. candidate, 1980-83, research supervised by D. McClure.

Joyce E. Anderson, Ph.D. candidate, 1981-83, research supervised by D. McClure.

Barry R. Davis, Ph.D. candidate, 1980-82, research supervised by S. Geman. Ph.D. requirements completed June 1982. Visiting Assistant Professor at Brown University, 1982-83.

Aytul Erdal, Ph.D. candidate, 1980-83, research supervised by S. Geman. Ph.D. requirements completed January 1983. Research Associate at Brown University, 1983.

Edmond Nadler, Ph.D. candidate, 1981-83, research supervised by D. McClure.

Charles Plumeri, Ph.D. candidate, 1980-81, research supervised by D. McClure. Ph.D. requirements completed June 1981.

Shoulamit Shwartz, Ph.D. candidate, 1980-83, research supervised by D. McClure. Thesis research compppleted June 1983. Degree will be awarded June 1984.

REFERENCES

1. Accomando, N., The implementation of maximum likelihood reconstruction algorithms for single photon emission tomography, Reports in Pattern Analysis No. 126, D.A.M., Brown University, 1982.

2. Agarwal, G. G. and Studden, W. J., Asymptotic integrated mean square error using least squares and bias minimizing splines, Ann. Statist. 8, 1307-1325, 1980.

3. Ahmad, I. A. and Lin, P. E., Nonparametric sequential estimation of a multiple regression function, Bull. Math. Statist. 17, 63-75, 1976.

4. Allen, D. M., The relationship between variable selection and data augmentation and a method for prediction, Technometrics 16, 125-127, 1974.

5. Bahadur, P. R., Rates of convergence of estimates and test statistics, Ann. Math. Statist. 38, 303-324, 1967.

6. Blum, J. and Walter, G., A simple solution to a non-parametric maximum likelihood estimation problem, Technical Report No. 43, Intercollege Division of Statistics, University of California, Davis, 1982.

7. Boneva, L. I., Kendall, D. and Stefanov, I., Spline transformations: Three new diagnostic aids for the statistical data-analyst, J. Roy. Statist. Soc. Ser. B 33, 1-71, 1971.

8. Budinger, T. F., Gullberg, G. T., and Huesman, R. H., Emission computed tomography, chapter 5 in Image Reconstruction from Projections: Implementation and Applications, G. T. Herman (ed.), Volume 32 of Topics in Applied Physics, Springer-Verlag, New York, 1979.

9. Chow, Y.-S , Estimation of Conformal Mappings, Ph.D. Dissertation, Division of Applied Mathematics, Brown University, 1980.

10. Chow, Y.-S., Geman, S. and Wu, L.-D., Consistent cross-validated density estimation, Ann. Statist. (to appear: March, 1983).

11. Devroye, L. P. and Wagner, T. J., Distribution-free consistency results in nonparametric discrimination and regression function estimation, Ann. Statist. 8, 231-239, 1980.

12. Duin, R. P. W., On the choice of smoothing parameters for Parzen estimators of probability density functions, IEEE Trans. on Computers C-25, 1175-1179, 1976.

13. Eastment, H. T. and Krzanowski, W. J., Cross-validatory choice of the number of components from a principal component analysis, Technometrics 24, 73-77, 1982.

14. Egerton, M. F. and Laycock, P. J., Some criticisms of stochastic shrinkage and ridge regression, with counterexamples, Technometrics 23, 155-159, 1981.

15. Freiberger, W. and Grenander, U., A triangulation sieve for restoring surface patterns, Reports in Pattern Analysis No. 104, D.A.M., Brown University, 1981.

16. Gasser, T. and Muller, H. G., Kernel estimation of regression functions, Smoothing Techniques for Curve Estimation, T. Gasser and M. Rosenblatt (eds.), Springer-Verlag, Berlin, 1979.

17. Geman, S., An application of the method of sieves: Functional estimator for the drift of a diffusion, Colloquia Mathematica Societatis Janos Bolyai 32, Nonparametric Statistical Inference, North-Holland, Budapest (Hungary), 1980.

18. Geman, S., Sieves for nonparametric estimation of densities and regressions, Reports in Pattern Analysis No. 99, D.A.M., Brown University, 1981.

19. Geman, S. and Hwang, C.-R., Nonparametric maximum likelihood estimation by the method of sieves, Ann. Statist. 10, 401-414, 1982.

20. Geman, S. and McClure, D. E., Characterization of a maximum-likelihood nonparametric density estimator of kernel type, Proceedings of the NASA Workshop on Density Estimation and Function Smoothing, L. F. Guseman, Jr. (ed.), Texas A&M University, College Station, 38-47, 1982.

21. Golub, G. H., Heath, M. and Wahba, G., Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21, 215-223, 1979.

22. Good, I. J. and Gaskins, R. A., Global nonparametric estimation of probability densities, Virginia J. Sci. 23, 171-193, 1972.

23. Gordon, L. and Olshen, R. A., Consistent nonparametric regression from recursive partitioning schemes, J. Mult. Analysis 10, 611-627, 1980.

24. Grenander, U., Abstract Inference, John Wiley & Sons, New York, 1981.

25. Grenander, U. and McClure, D. E., Guide to the APL Library of Mathematical Software, Division of Applied Mathematics, Brown University, Providence, RI (1979). Expanded editions (1980, 1981).

26. Habbema, J. D. F., Hermans, J. and van den Brock, K., Selection of variables in discriminant analysis by F-statistic and error rate, Technometrics 19, 487-493, 1977.

27. Helgason, S., The Radon Transform, Progress in Mathematics, Vol. 5, Birkhauser Boston, 1980.

28. Huesman, R. H., The effects of a finite number of projection angles and finite lateral sampling of projections on the propogation of statistical errors in transverse section reconstruction, Phys. Med. Biol. 22, 511-521, 1977.

29. Kronmal, R. and Tarter, M., The estimation of probability densities and cumulatives by Fourier series methods, J. Amer. Stat. Assoc. 63, 925-952, 1968.

30. Lawless, J. F., Mean squared error properties of generalized ridge estimators, J. Amer. Stat. Assoc. 76, 462-466, 1981.

31. Ludwig, D., The Radon transform on Euclidean space, Comm. Pure Appl. Math. XIX, 49-81, 1972

32. McClure, D. E., Nonlinear segmented function approximation and analysis of line patterns, Quart. Appl. Math. 33 (1975), 1-37.

33. McClure, D. E., Image models in pattern theory, Computer Graphics Image Processing 12, 309-325, 1980. Republished in Image Modeling, A. Rosenfeld (ed.), Academic Press, New York (1981), 259-275.

34. McClure, D. E., Estimation of planar sets from Poisson projections, Proceedings of the NASA Workshop on Density Estimation and Function Smoothing, L. F. Guseman, Jr. (ed.), Texas A&M University, 32-37, 1982.

35. McClure, D. E., Estimating the intensity function of a planar Poisson process from projection data, manuscript in preparation. Invited paper for special session on "Inference for Stochastic Processes", Annual meeting of the Institute of Mathematical Statistics, Cincinnati, OH, 1982.

36. Mistretta, C. A., Crummy, A. B., and Strother, C. M., Digital Angiography: A Perspective, Diagnostic Radiology 139, 273-276, 1981.

37. Nguyen, H. T. and Pham T. D., Indentification of nonstationary diffusion model by the method of sieves, Siam J. Control and Optimization (to appear: 1982).

38. Prakasa Rao, B. L. S., Estimation of a unimodal density, Sankhya Ser. A 31, 23-36, 1969.

39. Pratt, W. K., Digital Image Processing, John Wiley & Sons, New York, 1978.

40. Sacks, J. and Ylvisaker, D., Design for regression problems with correlated errors III, Ann. Math. Statist. 41, 2057-2074, 1970.

41. Schoenberg, I. J., Spline functions and the problem of graduation, Proc. Nat. Acad. Sci. 52, 947-950, 1964.

42. Schuster, E. F. and Gregory, G. G., On the nonconsistency of maximum likelihood nonparametric density estimators, Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, W. F. Eddy (ed.), Springer-Verlag, New York, 295-298, 1981.

43. Schuster, E. and Yakowitz, S., Contributions to the theory of nonparametric regression, with application to system identification, Ann. Statist. 7, 139-149, 1979.

44. Scott, D. W. and Factor, L. E., Monte Carlo study of three data-based nonparametric probability density estimators, J. Amer. Statist. Assoc. 76, 9-15, 1981.

45. Shepp, L. A. and Vardi, Y., Maximum likelihood reconstruction for emission tomography, Bell Laboratories preprint, 1982.

46. Silverman, B. W., Choosing the window width when estimating a density, Biometrika 65, 1-11, 1978.

47. Speckman, P., Efficient nonparametric regression with cross-validated smoothing splines, preprint, 1982.

48. Spiegelman, C. and Sacks, J., Consistent window estimation in nonparametric regression, Ann. Statist. 8, 240-246, 1980.

49. Stone, M., Cross-validation and multinomial prediction, Biometrika 61, 509-515, 1974.

50. Stone, M., Cross-validation: A review, Math. Oper. Sch. Statist., Ser. Statistics 9, 127-139, 1978.

51. Tapia, R. A. and Thompson, J. R., Nonparametric Probability Density Estimation, Johns Hopkins University Press, Baltimore, MD, 1978.

52. Thomasset, F., Implementation of Finite Element Methods for Navier-Stokes Equations, Springer Series in Computational Physics, Springer-Verlag, New York, 1981.

53. Utreras, F., Cross-validation techniques for smoothing spline functions in one or two dimensions, Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics 757. T. Gasser and M. Rosenblatt (eds.), Springer-Verlag, Berlin, 1979.

54. Wahba, G., Data-based optimal smoothing of orthogonal series density estimates, Ann. Statist. 9, 146-156, 1981.

55. Wahba, G. and Wold, S., A completely automatic French curve: Fittin g spline functions by cross-validation, Comm. Statist. 4, 1-17, 1975.

56. Wald, A., Note on the consistency of the maximum likelihood estimate, Ann. Math. Statist. 20, 595-601, 1949.

57. Wegman, E. J., Maximum likelihood estimation of a unimodal density, II, Ann. Math. Statist. 41, 2169-2174, 1970.

58. Wong, W. H., Expected information criterion for the smoothing parameter of density estimates: an elucidation of the modified likelihood. Technical Report No. 589, Dept. of Statistics, Univ. of Wisconsin, Madison, Wisconsin, 1979.

59. Wong, W. H., On the consistency of cross-validation in kernel nonparametric regression, Technical Report No. 139, Department of Statistics, University of Chicago, 1982.